US 20020168664A1

(19) United States

(12) Patent Application Publication (10) Pub. No.: US 2002/0168664 A1

Murray et al. (43) Pub. Date: Nov. 14, 2002

(54) AUTOMATED PATHWAY RECOGNITION SYSTEM

(76) Inventors: Joseph Murray, Berkeley, CA (US); Donna Hendrix, Berkeley, CA (US); Daniel J. Chin, Foster City, CA (US)

Correspondence Address:
BOZICEVIC, FIELD & FRANCIS LLP
200 MIDDLEFIELD RD
SUITE 200
MENLO PARK, CA 94025 (US)

(21) Appl. No.: 10/090,698

(22) Filed: Mar. 4, 2002

Related U.S. Application Data

(63) Continuation-in-part of application No. 09/365,587, filed on Jul. 30, 1999. Continuation-in-part of application No. PCT/US00/20603, filed on Jul. 28, 2000.

Publication Classification

(51) Int. Cl.[7] C12Q 1/68; G06F 19/00; G01N 33/48; G01N 33/50

(52) U.S. Cl. 435/6; 702/20

(57) ABSTRACT

There is a pressing need for computer-implemented tools that can summarize and present the enormous amounts of public literature to facilitate analysis of gene expression data. The present invention provides techniques and systems for efficiently integrating public literature regarding gene function with data from gene expression profiling experiments. Information from literature databases relating to a particular set of DNA sequences of known expression pattern is retrieved, processed, cross-referenced and viewed to provide further information about a particular DNA sequence to facilitate its identification as a candidate gene.

PGPUB-DOCUMENT-NUMBER: 20020168664

PGPUB-FILING-TYPE: new

DOCUMENT-IDENTIFIER: US 20020168664 A1

TITLE: Automated pathway recognition system

PUBLICATION-DATE: November 14, 2002

INVENTOR-INFORMATION:

NAME	CITY	STATE	COUNTRY
RULE-47			
Murray, Joseph	Berkeley	CA	US
Hendrix, Donna	Berkeley	CA	US
Chin, Daniel J.	Foster City	CA	US

US-CL-CURRENT: 435/6, 702/20

ABSTRACT:

There is a pressing need for computer-implemented tools that can summarize and present the enormous amounts of public literature to facilitate analysis of gene expression data. The present invention provides techniques and systems for efficiently integrating public literature regarding gene function with data from gene expression profiling experiments. Information from literature databases relating to a particular set of DNA sequences of known expression pattern is retrieved, processed, cross-referenced and viewed to provide further information about a particular DNA sequence to facilitate its identification as a candidate gene.

---------- KWIC ---------

Summary of Invention Paragraph - BSTX (6):

[0006] More recently, gene expression profiles have been examined using methods that can cross-compare the expression profiles of many thousands of genes across many different experiments (for example Eisen et al P.N.A.S. 95, 14863-8). These methods employ pattern recognition algorithms to cluster genes with a similar expression patterns facilitating the facile identification of groups of genes that are co-regulated. Both supervised and unsupervised pattern recognition algorithms can be used to for clustering. Supervised pattern recognition algorithms require a priori knowledge that forms a training set, whereas unsupervised pattern recognition algorithms do not need a priori knowledge and are typically used to discover latent patterns. Many **unsupervised clustering** methods have been applied to gene expression profile data: these include hierarchical, K-means, self-organizing maps (Tamayo et al. PNAS 96:2907-12), or support vector machines (M. Brown et al. PNAS 97:262-7).

Summary of Invention Paragraph - BSTX (13):

[0012] In another embodiment, the present invention provides a method for analyzing a group of genes identified through analysis of gene expression profiling experiments, wherein the groups of genes have been grouped together by a commonality in their gene expression patterns. Clustering algorithms may be employed to automatically group genes by their expression pattern and a cluster of genes may represent a group of genes. These clustering algorithms may be supervised or unsupervised. A further embodiment of the invention

provides a method for using both supervised and **unsupervised clustering** algorithms to automatically group genes by their expression pattern. The gene expression data analyzed may be from microarray experiments.

Detail Description Paragraph - DETX (81):

[0110] FIG. 8 depicts the gene expression profile data stored in the database according to an embodiment of the present invention. The four tables depicted in FIG. 8 correspond to a summary of the array result conditions ("ArrayResults" 240), the summarized array data ("ArrayData" 250), the details of the probe(s) ("Probe" 260), and the **raw data** ("RawData" 270). The array result conditions table 240 contains attributes that describe a unique experimental identifier ("ExptlD" 240-a), the corresponding bar code ("BarCode" 240-b), the link for probe 1 ("Probe1" 240-c), the link for probe 2 ("Probe2" 240-d), a term that describes the grid pattern ("GridPattern" 240-e), the clone set identifier ("CloneSet" 240-f), the link to array data ("ArrayData" 240-g), and a comment ("Comment" 240-h). The array data table 250 contains attributes to describe the experimental identifier ("ExptlD" 250-a), the name of the cDNA sequence ("seqFile" 250-b), the arithmetic mean of the background or normalized data ("Mean" 250-c), the standard deviation ("StdDev" 250-d), the ratio of any paired means derived from simultaneous application of two probes ("Ratio" 250-e), the time point at which the probes were made ("TimePt" 250-g), the biological state (e.g. diseased or normal) of the probe's mRNA origin ("State" 250-h), the clustering method ("ClusterMethod" 250-i), the cluster number ("Cluster" 250-j), the total number of clusters ("TotalClusters" 250-k), the cluster order pattern derived from the auto-regression analysis used in the causality analysis ("ClusterOrder" 250-l) and the date of the clustering ("ClusterDate" 250-m). Other attributes, such as patterns arising from ANOVA analysis or other parametric or non-parametric tests, and/or propagated error values may be added.

Detail Description Paragraph - DETX (83):

[0112] The **raw data** table 270 contains attributes for the experimental identifier ("ExptlD" 270-a), the sequence name ("seqFile" 270-b), the probe name ("Probe" 270-c), the raw intensity value ("RawValue" 270-d), the local background or normalization factor ("LocalBgnd/factor" 270-e), and the arithmetically corrected intensity value ("CorrectedValue" 270-f).

Claims Text - CLTX (7):

6. The computer-implemented method of claim 5, wherein said clustering is **unsupervised clustering**.

Claims Text - CLTX (9):

8. The computer-implemented method of claim 5, wherein said clustering is a combination of supervised and **unsupervised clustering**.